Math 473 HW 3 Spring 2023. Due Friday, Feb. 10.

1) 1.13: The data set consists of information from 927 1st born children to mothers who chose to breast feed their child. The event was time in weeks until weaned (instead of death). Complete the following table used to produce the lifetable estimator (on a separate sheet of paper).

I_j	d_j	c_j	n_j	n'_j	$\frac{n'_j - d_j}{n'_j}$	$\hat{S}_L(t)$
$[0, 2)$	77	2	927	926	0.9168	1.0000
$[2, 3)$	71	3	848	846.5	0.9161	0.9168
$[3, 5)$	119	6	774	771	0.8457	0.8399
$[5, 7)$	75	9	649	644.5	0.8836	0.7103
$[7, 11)$	109	7	565	561.5	0.8059	0.6276
$[11, 17)$	148	5	449	446.5	0.6685	0.5058
$[17, 25)$	107	3	296			0.3381
$[25, 37)$	74	0	186			
$[37, 53)$	85	0	112			
$[53, \infty)$	27	0	27			

2) 1.21: This problem examines the myelomatosis data (a cancer causing tumors in the bone marrow) with SAS using the Kaplan Meier product limit estimator. Obtain the SAS program for 1.21 from (http://parker.ad.siu.edu/Olive/survhw.txt). Obtain the output from the program in the same manner as 1) through 4) from the SAS handout.

a) The output should be roughly 3 pages and a graph. Include this output in *Word*.

b) From the summary statistics of the first page of output, about when do 50% of the patients die?

c) From the first page of output (perhaps), what is the 95% CI for the time when 50% of the patients die?

d) From the 3rd page of output (perhaps), what is the 95% CI for $S_Y(13)$. This is the log log transformed CI, so will differ from the CI in e).

e) Make the CI using $\hat{S}_K(13)$ and $SE(\hat{S}_K(13))$ obtained from the 1st page of output (perhaps). If the interval is (L, U), use $(\max(0, L), \min(U, 1))$ as the final interval.

f) From the plot of $\hat{S}_K(t)$ for the KM estimator, briefly explain survival for days 0–250 and for days 250–2250.

3) 1.23: The length of times of remission (time until relapse) in acute myelogeneous leukemia under maintenance chemotherapy for 11 patients is
$9, 13, 13+, 18, 23, 28+, 31, 34, 45+, 48, 161+$.

a) Following the example done in class and 12) from the exam 1 review, make a table with headers $t_{(j)}$, γ_j, t_i, n_i, d_i and $\hat{S}_K(t_i)$. Then compute the Kaplan Meier estimator. (You can check it with the R output obtained in b).)

b) Following the R handout, get into R. Copy and paste 1.23 commands from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into R. Hit `Enter` and a plot should appear. Copy and paste the R output with header (time ... upper 95% CI) into *Word*. Following the R handout, click on the plot and hold down the *Ctrl* and c buttons simultaneously. Then in the *Word* Edit menu, select "paste."

Include this output with the homework. The center step function is the Kaplan Meier estimator $\hat{S}_K(t)$ while the lower and upper limits correspond to the confidence interval for $S_Y(t)$.

c) Write down the 95% CI for $S_Y(23)$ and then verify the CI by computing $\hat{S}_K(23) \pm 1.96SE(\hat{S}_K(23))$.

4) 1.24: Copy and paste commands for 1.24a) and 1.24b) from (http://parker.ad.siu.edu/Olive/survhw.txt) for this problem into R.

The commands make the KM estimator for censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. The KM estimator attempts to estimate $S_Y(t) = \exp(-t)$. The points in the plot are $S_Y(t_{(j)}) = \exp(-t_{(j)})$, and the points should be within the confidence intervals roughly 95% of the time (actually, if you make many plots the points should be in the intervals about 95% of the time, but for a given plot you could get a "bad data set" and then the rather more than 5% of the points are outside of the intervals).

a) Copy and paste the commands for 1.24a) and hit `Enter`. Then copy and paste the plot into *Word*.

b) Copy and paste the commands for 1.24b) and hit `Enter`. Then copy and paste the plot into *Word*.

c) As the sample size increases from $n = 20$ to $n = 200$, the CIs should become more narrow. Can you see this in the two plots? Are about 95% of the plotted points within the CIs? If it is hard to estimate the percentage, say so.

5) 1.25: Copy and paste the commands for problem 1.25 into R. a) Type the command `kmsim2(n=10)`, hit `Enter` and include the output in *Word*.

This program computes censored data $T = \min(Y, Z)$ where $Y \sim EXP(1)$. Then a 95% CI is made for $S_Y(t_{(j)})$ for each of the $n = 10$ $t_{(j)}$. This is done for 100 data sets and the program counts how many times the CI contains $S_Y(t_{(j)}) = \exp(-t_{(j)})$. The scaled lengths are also computed. The ccov is the count for the classical $\hat{S} \pm 1.96SE(\hat{S})$ interval while p4cov is for the plus 4 CI. The lcov is based on a CI that uses $\log(\hat{S})$ and llcov is based on a CI that uses $\log(-\log(\hat{S}))$. The 1st 3 CIs are not made if the last case is censored so NA is given. The plus 4 CI seems to be good at $t_{(1)}$ and $t_{(n)}$.